Fermilab

Title: Understanding and Coping with Hardware and
Software Failures in a very Large Trigger Farm

Time critical event filtering applications running on trigger farms with thousands of processors will be subject to a large number of failures within the software and hardware systems. A system with these properties is the one proposed for BTeV, a proton/antiproton collider experiment at Fermi National Accelerator Laboratory. It is easily imaginable that the number of fault types, their relationships, and rate at which they will occur will be sufficiently great that there will be a large impact on the effectiveness of the trigger system. It is likely that an administrative staff and cast of experiment operators will not be able to service simple problems or analyze complex problems in a timely fashion to avoid data loss. The RTES (Real Time Embedded Systems) collaboration is a group of physicists, engineers, and computer scientists working to address the problem of reliability in large scale clusters with real-time constraints such as this. RTES is defining software infrastructure to detect, diagnose, and recover from errors not only at the system administrative level, but also at the application level. This infrastructure must be highly scalable (to minimize bottlenecks or single points of failure), verifiable (does what it is supposed to in a timely fashion), extendible by users (grow new detection/analysis methods as they are discovered), and dynamic (changeable as it is operating). The problem is being approached using a hierarchy of monitoring and control elements, with capabilities and decision making ability becoming greater as one moves away from source. At the top are system modeling and evaluation tools.

-James Kowalkowski